MD Concept: A Model for Integrating Medical Knowledge

Yves Lévesque MD*◊; A. Robert LeBlanc, PhD◊; Michel Maksud, MSc*
◊ Institut de Génie Biomédical, Université de Montréal, Montréal, CANADA
* Développement Purkinje Inc., 7333 place des Roseraies, suite 401,
Anjou, Qc, CANADA, H1M 2X6

ABSTRACT

Many integrated clinical information systems depend on large knowledge bases containing dictionary of terms as well as specific information about each term and the relationships between terms. We propose a knowledge base model called MD Concept which is based on a semantic network and uses an object-oriented paradigm and relational tables. A prototype has been developed which integrates the Unified Medical Language System (UMLS) with other databases including the Systematized Nomenclature of Medicine (SNOMED II), the Diagnostic and Statistical Manual of Mental Disorders (DSM-IIIR) and a pharmaceutical database. We demonstrate how a user can easily navigate in this knowledge world using a browser .

INTRODUCTION

Health professionals deal with information: They need access to knowledge in the medical field [1], to data about particular patients (medical records) and they need to make links between the two [2]. To be more useful, all these information systems should be integrated [3,4]. Terminology is a key factor for integrating these systems [5]. Many terminologies are currently in use but none is adequate for all purposes [6,7]. Many of these terminologies, in particular the *Systematized Nomenclature of Medicine* (SNOMED III) [8] and the *Unified Medical Language System* (UMLS) [7], have evolved from simple lists of terms and codes to knowledge representation systems. Such systems make it possible to integrate clinical tools such as computerized medical record systems, decision support systems, and information retrieval systems. They could help information systems to show "intelligent" behavior.

We propose a model for integrating this knowledge in a comprehensive knowledge base called MD Concept. A prototype has been developed to explore the model. UMLS Knowledge Sources [9] were used as the core of the knowledge base.

OBJECTIVES

Ideally, a knowledge base such as MD Concept should contain a multilingual terminology covering all of medicine along with information on the terms and concepts represented, for example synonyms, translations, the codes for various other systems of classification and nomenclature (*eg* CPT), definitions, hierarchical (taxonomic) and other semantic information, linguistic (lexical, syntactic) information, and so on. There should also be data specific to the type or class of the term; for example, along with a particular commercial drug should be stored information about its ingredients and their concentrations; indications and contraindications; side effects; manufacturer; dispensing information, and so on.

A user should be able to consult the knowledge base to get answers to questions such as: What is Cushing's disease? What are its symptoms or its treatment? What is its ICD-9CM code? Another question might be, what diseases affecting the meninges are caused by a virus?

A knowledge base should confer some "intelligence" on information systems. For example, if a clinician were to ask the computerized medical record if his patient has heart disease, the system should know that myocardial infarction *is a* heart disease so as to report that the patient has a myocardial infarction. Linguistic information can also enable the system to understand queries and generate reports in a more "natural" language.

THE MODEL

We use the ANSI/SPARK Model [10] to describe our knowledge base architecture. This architecture is divided in three levels: the external level (user view), the conceptual level (application view), and internal level (physical model).

The External Level: semantic network

The external level is the *user view* of the knowledge base, how the user understands it and interacts with it. Following Gabrieli [11] and Rector, Nowlan and Kay [12], our model can be represented as a *semantic network* [13] in which elements of information (nodes) are linked together by relations (arcs).

Elements, or nodes, include words or groups of words, texts, numbers, images, sounds, etc. Most of the elements in MD Concept represent *concepts*. A concept is an abstraction; it is something which has meaning all by itself. For example, symptoms such as earache, diseases such as measles, treatments, lab tests, etc. are all concepts. As in the UMLS model [14], a concept typically will be represented by several different synonyms, or *terms*, possibly in more than one language. A *term* is defined as a word or a group of words representing a concept. As

0195-4210/94/$5.00 © 1994 AMIA, Inc.

examples, "Sexually transmitted disease", "STD", "Venereal disease", "Maladie transmissible sexuellement", "Maladie vénérienne" are the terms representing the concept of a disease that is transmitted by sexual contact. Each concept has a *preferred term* in each language.

A term can have a number of *lexical variations* (called *strings* in UMLS) that are minor variations of that term caused by singular-plural forms, order of words, etc. "Venereal Diseases", "Venereal, disease", "Disease, venereal ", "Diseases, venereal" are lexical variations of the same term. Each term has a *preferred lexical variation.*

In our model, the above semantic structure is actualized by assigning each element to a *class*, such as CONCEPT, TERM, LEXICAL VARIATION, CODE, etc. As in the UMLS, each element in the class CONCEPT is further subclassified into one or more *semantic types,* such as DISEASE, BACTERIA, HORMONE, TISSUE, etc. For example, the concept "Insulin" has two semantic types: HORMONE and PHARMACOLOGICAL SUBSTANCE. It is important to note that semantic types are themselves concepts.

A **relation** is a link between two elements (*table 1*). Relations are bidirectional: if myocardial infarction *has as a symptom* chest pain, then chest pain *is a symptom of* myocardial infarction. Most relations are heritable: if myocardial infarction *is a* heart disease and a heart disease *has as a site* heart, then we can conclude, if nothing else is specified, that myocardial infarction *has as a site* the heart.

The types of relations which can exist between any two concepts are determined by the semantic types of the two concepts; for example, "BACTERIA *causes* DISEASE" is possible, but "BACTERIA *causes* HORMONE" is not possible.

The Conceptual Level: the Object Oriented Model

The conceptual level represents knowledge as viewed by software applications. We use the object-oriented model [15]. Elements correspond to *objects* of a class and relations correspond to *methods* (sub-programs or functions) to find related elements.

There is an object class for each class of elements: TERM, CONCEPT, LEXICAL VARIATION, CODE, etc. These classes derive from the virtual class ELEMENT. Semantic type concepts derive from the CONCEPT class.

All elements have a *label* (a name). An element of the CONCEPT class has as its label the PREFERRED TERM for that concept. An element of the TERM class has the PREFERRED LEXICAL VARIATION as its label. For the LEXICAL VARIATION class, the label is simply the words for the lexical variation.

ELEMENT	RELATION	ELEMENT
myocardial infarction	is a	heart disease
myocardial infarction	ICD-9 code	410
myocardial infarction	has symptom	chest pain
chest pain	is a symptom of	myocardial infarction
venereal disease	is synonymous with	sexually transmitted disease

Table 1: Examples of elements and relations

For the CODE class, a concatenation of the source and the code (*eg* "ICD-9CM 412.34") forms the label.

Each class has its own data members (attributes): elements from the CONCEPT class have an ID, a preferred French term, a preferred English term, one or more semantic types, a syntactic category, etc. Elements from the TERM class have an identifier (ID), a language, a concept, a lexical tag, etc.

An object class includes *methods*, or programs, which define the relations between elements of that class and elements belonging to other classes. Methods can be quite simple, *eg* table lookup, but can also be very complex, including the capacity to make deductions. For example, suppose we want to know the site for the element "Myocardial infarction" (class CONCEPT). We would activate the method *has as a site* to search for a related element. If the method cannot find an element using this relation, it would next try to find elements using the relation *is a.* If it finds any (in our example, the method would come up with "heart disease"), it then recursively searches for the relation *has as a site* (it would find the element "heart"). The method can then infer that "myocardial infarction" *has as a site* "the heart".

This deduction of new knowledge uses primarily *is a* relations, but in some cases, the *is part of* relation provides more information. For example, to find diseases that "*have as a site* the heart", one should also look for diseases that *have as a site* part of the heart (myocardium, left ventricle, etc.)

Knowledge bases differ from simple data bases in this capacity to infer new information [16]. However, the user must be told that the knowledge was obtained by inference and how it was inferred.

As defined in the object oriented model, methods can be *overloaded.* So, in a class derived from a superclass, the method of the derived class can replace the method of the superclass. See [15] for a more complete description of the object oriented model.

The Internal Level: the Relational Model

The internal level represents how the knowledge base is physically stored in files. The relational model [10] is well suited to this task. Information is stored into and retrieved from relational tables using methods.

THE PROTOTYPE

A prototype was developed to explore the model. Borland C++ 3.1 and Object Window Library (OWL 1.0) were used to create a Windows 3.1 application running on an Intel486 DX2/66 microcomputer. The database is accessed via IS, a locally developed library of programs for managing indexed sequential files. Two modules have been implemented to date: MD Concept Integrator and MD Concept Browser.

MD Concept Integrator

The integrator can input files from multiple sources and output an integrated relational database. For the prototype, we integrated databases from UMLS, SNOMED II [17], DSM-IIIR [18], and a pharmaceutical database derived from the Canadian Drug Identification Code and monographs from the University of British Columbia Drug and Poison Information Center [19].

The process was as follows: first, using UMLS Knowledge Sources (meta 1.3, April 1993 [9]), the integrator set up relational tables containing 152,444 concepts, 202,000 terms, 279,238 lexical variations, 680,345 semantic relations, 52,085 concept definitions, 311,046 codes, and many of their attributes. Our data base differs from UMLS in that lexical variations, terms and concepts are placed into separate normalized tables .

Second, using SNOMED codes already present in UMLS and the SNOMED II bilingual database, we added 11,814 SNOMED II French terms and their lexical variations. We used the SNOMED code (*termcode)* and *the* English term (*enomen)* to find an English lexical variation. We tried to find an existing French term and lexical variation equivalent to the SNOMED French term *(fnomen).* If we did not find one, we added the term and/or the lexical variation. We did not add new concept. Translation relations were also created between French and English terms. The SNOMED *reference* field was used to add semantic relations (*has location, is associated with, cause*, etc) between concepts. Employing similar methods, we added french terms and translation relations from a bilingual DSM-IIIR database.

Finally, from the pharmacological database, 5,432 commercial drug names with their ingredients, codes, dose, manufacturer, form and route of administration were added as new concepts (of the new semantic type "Commercial Medication"), terms and lexical variations. Relations to 336 medication monographs and 14,318 pharmacological interactions between medications were integrated. New relations *have ingredient, made by, form, route, monograph, interact with* were created to link these commercial medications into the UMLS and SNOMED data.

The complete MD Concept knowledge base consists of approximately 230 megabytes of relational data files.

MD Concept Browser

The browser allows a user to find a particular element in the semantic network and to browse from one element to another using relations. There are two ways to access an element: with key words or with a code (Fig. 1), either gives rise to a lexical variation. The lexical variation can be used to access the term for the concept (*ie* the concept itself). From here, relations can be used to find the related concept, other terms (synonyms), terms in other languages, and all their lexical variations. For any term, one can look up codes (*e.g.* DSM-IIIR, SNOMED, ICD-9CM). It is possible to navigate to other concepts using any of the semantic relations defined for that concept.

Some relations are actually combinations of several relations. For example, as seen in Fig. 1, the user can go from a keyword directly to the concept (keyword →lexical variation→term→concept). If the *codes* relation for a concept are requested, all codes for all lexical variations for all terms for that concept will be found.

A simple generic *Elements-Relations-Messages* dialog box (Fig. 2) is used for browsing. It contains two lists and a message field. The *Element List* shows the element labels for any class. When one element is selected, the *Relation List* shows all relations defined for the class and the semantic type of that element. When the user selects one relation, the browser shows the new elements either in the *Message Field* (for a simple text element) or in a new overlapping *Elements-Relations-Messages* dialog box. The M*essage Field* is also used to show a definition for the relation or the path that MD Concept used to infer some relations (as in the "Myocardial infarction *has as a site* heart" example). The caption on each

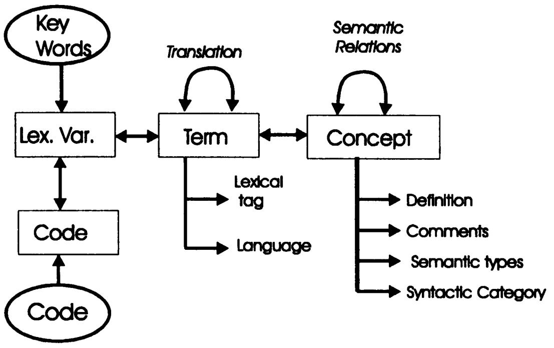
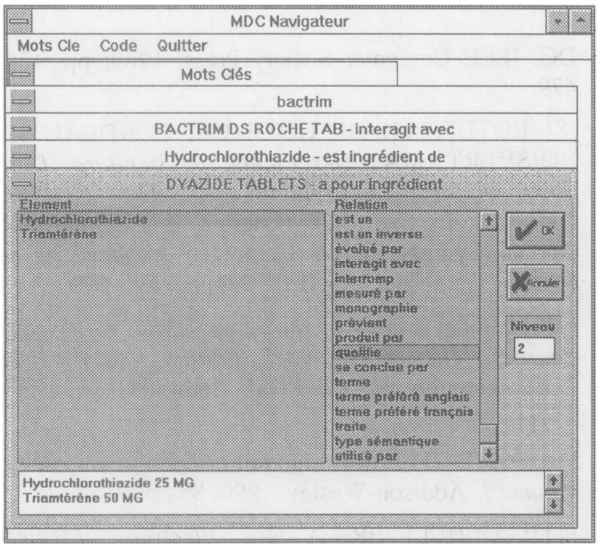


Fig 1: Browsing in the semantic network


MDC Navigateur
Mots Cle
Code
Quitter
Mots Clés
bactrim
BACTRIM DS ROCHE TAB - interagit avec
Hydrochlorothiazide - est ingrédient de
DYAZIDE TABLETS - a pour ingrédient
Element
Relation
OK
Annuler

Element
Hydrochlorothiazide
Triamtérène

Relation
est un
est un inverse
évalué par
interagit avec
interromp
mesuré par
monographie
prévient
produit par
qualifie
se conclue par
terme
terme préféré anglais
terme préféré français
traite
type sémantique
utilisé par

Niveau 2

Hydrochlorothiazide 25 MG
Triamtérène 50 MG

Fig. 2: MD Concept browser with overlappec *Elements-Relations-Messages* dialog boxes

overlapping dialog box shows the elements and the relations chosen by the user. A specialized dialog box displays the monograph text for pharmacological ingredients of a medication.

Fig. 2 shows a series of overlapped *Elements-Relations-Messages* dialog boxes. The caption bars of each window show the path followed by the user: starting with the keyword *bactrim,* MD Concept found 5 concepts (*ie* it found 5 lexical variations containing the word bactrim); then the terms for each of these variations, and finally the concepts for these terms (timing: less than 2 seconds).

The next caption bar shows that the user then selected the element *"BACTRIM DS ROCHE TAB"* and the relation *"interagit avec" (interacts with).* The system located 21 elements. Out of these, the user selected element *"Hydrochlorothiazide"* and the relation *"est ingrédient de" (is ingredient of)* to get 43 commercial medications (timing: 2 seconds). Finally, the element *"DYAZIDE TABLETS"* and the relation *"a pour ingredient" (has ingredient)were chosen.* The topmost Element List shows the two ingredients of *DYAZIDE TABLET* (timing: less than 1 second). The Message Field shows concentrations for these ingredients.

In summary: *bactrim is a keyword of the concept BACTRIM DS ROCHE TAB that interacts with Hydrochlorothyazide that is an ingredient of DIAZIDE TABLETS that has for ingredient Hydrochlorothiazide 25 MG and Triamterene 50 MG.*

The relational model gives fast access to knowledge. The response time was surprisingly good, considering the size of the database (230 megabytes): 1.1 million words in 300,000 lexical variations, hundreds of thousands of relations, etc.

DISCUSSION

MD Concept extends the UMLS model in three ways: with a simplified user interface for browsing in the knowledge base, with a mechanism for inferring new knowledge, and with additional content.

At the user level (external level), a semantic network appears to be a simple but effective way to represent knowledge. Compared to other browsers like *Metacard* or *COACH* [9] used with the UMLS, MD Concept uses a generic *"Elements-Relations-Messages"* dialog box to navigate in his knowledge base. While this generic dialog box is simple to use and is adequate for general browsing, a series of specialized user interfaces would be useful for regular use of the knowledge base in certain domains.

The methods of the object model facilitate the representation of inheritance of knowledge and the inference of new knowledge. As this inference process can lead to erroneous conclusions (e.g. with ICD-9 codes), the user must always be told how the information was inferred.

We found that it is possible for a personal computer to handle a large knowledge base.

We added some content to the UMLS: french terms with diacritical marks and upper-lower case; new, clinically useful semantic relations; and a pharmaceutical database. The inclusion of medication data demonstrates that MD Concept can be more than a terminological knowledge base and that different types of knowledge can be added using all the codes and lexical variations provided by the UMLS.

The UMLS was of great help in this project. Because it integrates several knowledge sources it formed an ideal core for MD Concept. Unfortunately, the UMLS is still an experimental project and it contains many inconsistencies; for example, many *is a* relations are defined as *unspecified hierarchical relations* and many clinically important semantic relations are missing. SNOMED II was useful in adding these important semantic relations, in particular *location of* and *cause*. But these relations (*reference* field) are not always explicit between elements of different axes. SNOMED also provides many important French terms with diacritical marks. The UMLS includes many French terms but they are in capital letters without marks.

CONCLUSIONS

MD Concept is a prototype and its content has not been tested extensively. Our objective was to evaluate the possibility of integrating existing knowledge sources into a large usable knowledge base. We found that:

- semantic networks and a generic *Elements-Relations-Messages* dialog box can be used to represent and browse into a wide variety of knowledge;
- methods are useful for implementing knowledge inheritance;
- the UMLS is useful as the core of a terminological knowledge base and that other type of knowledge can be added to the UMLS;
- a large knowledge base containing hundreds of thousands of terms and much additional information can be implemented on a microcomputer.

REFERENCES

[1] LEAO BdF, MANTOVANI, ROSSI FRI, ZIELINSKY P. *Incorporating knowledge to databases - a solution to complex domains.* in: Frise M, ed., Proceedings of the 16th annual symposium on computer application in medical care. Washington DC: McGraw Hill, 1992; pp. 234-238

[2] WEED LL et al. *Representation of medical knowledge and PROMIS* in Proceedings of the second Annual Symposium in Computer Applications in Medical Care, 1978; pp. 368-400.

[3] LINNARSSON R, WIGERTZ, O. *The Data Dictionary- A controlled Vocabulary for Integrating Clinical Databases and Medical Knowledge Bases.* Methods of information in medicine 28(1989); pp. 78-85.

[4] TIMMERS T, van MULLIGEN, EM, van den HEUVEL F. *Integration of an Object Knowledge Base into a Medical Workstation.* in: Clayton P.D., ed., Proceedings of the 15th annual symposium on computer application in medical care. New York: McGraw Hill, 1991; pp. 654-658.

[5] RECTOR AL, NOWLAN WA, KAY S. *Conceptual Knowledge: the core of medical information systems.* in LUN, K.C. et al, ed., MEDINFO 92. North-Holland: Elsevier Science Publishers; 1992; pp. 1420-1426.

[6] CIMINO JJ, BARNET GO. *Automated translation between Medical Terminologies using Semantic Definitions.* MD Computing Vol. 7 No 2 1990; pp. 104-109.

[7] HUMPREYS BL. *Building the Unified Medical Language System.* in: Kingsland LC III, ed. Proceedings of the thirteenth annual symposium on computer application in medical care. Washington DC: IEEE Computer Society Press, 1989; pp. 475-479.

[8] ROTHWELL DJ, COTE RA, CORDEAU JP, BOISVERT MA. *Developing a Standard Data Structure for Medical Language - The SNOMED Proposal.* In SAFRAN, Charles, ed. Seventeenth Annual Symposium on Computer Applications in Medical Care. McGraw Hill 1994; pp. 695-699.

[9] National Library of Medicine. *UMLS Knowledge Sources, 4th experimental Edition- April 1993 Documentation and CD-ROM.* Bethesda MA, 1993; 157p.

[10] DATE CJ. *An Introduction to Database Systems, volume 1.* Addison-Wesley, 1990; 854 p.

[11] GABRIELI ER. *A New Electronic Medical Nomenclature.* Journal of medical systems. Vol. 13 No 6 1989; pp. 355-373.

[12] RECTOR AL, NOWLAN WA, KAY S. *Unifying medical information using an architecture based on descriptions.* in: Miller, Randolf A., ed. Proceedings of the fourteenth annual symposium on computer application in medical care. Los Alamitos, CA: IEEE Computer Society Press, 1990; pp. 190-194.

[13] SOWA JF. *Semantic Networks.* in SHAPIRO, Stuart C. Encyclopedia of Artificial Intelligence. New York, Wiley, 1992; p 1493-1511

[14] TUTTLE MS, SPERZEL WD, OLSON NE, ERLBAUM, MS, et al. *The Homogenisation of the Metathesaurus Schema and Distribution Format,* in: Frise M, ed., Proceedings of the 16th annual symposium on computer application in medical care. Washington DC: McGraw Hill, 1992; pp. 299-303

[15] BOOCH G. *Object Oriented Design with Applications.* Redwood City, CA. Benjamin/Cummings 1991; 580p.

[16] TOURETZKY DS. *Inheritance Hierarchy* in SHAPIRO, Stuart C. Encyclopedia of Artificial Intelligence. New York, Wiley, 1992; pp. 690-701.

[17] COTE RA. *SNOMED: Systematized Nomenclature of Medicine.* Second edition. College of American Pathologist. Skokie, Il. 1979- 1982.

[18] American Psychiatric Association. *Diagnostic and Statistical Manual of mental Disorders.* Third Edition, Revised (DSM-III-R). Washington D.C. American Psychiatric Association, 1987.

[19] LEATHEM AM, CADARIO BJ. *Drug Information Reference* third edition. Vancouver B.C. Drug And Poison Information Center.1993. 1518 p.